INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multi-resolution approach to Identification of spoken languages and to improve overall Language Diarization System using Whisper Model

Bhavik Vachhani[1],*Dipesh Singh*[2], *Rustom Lawyer*[3]

[1,2,3]Augnito India Private Limited, India

bhavik.v@augnito.ai, dipesh.singh@augnito.ai, rustom.lawyer@augnito.ai

Abstract

This research paper investigates the effectiveness of the Whisper decoder for Language Identification (LI) and Language Diarization (LD) tasks. An audio accent detection system was used as an attention mechanism to narrow down the Whisper LI output classes. The LI system was tested on different audio resolutions ranging from 1.0 to 11.0 seconds, and the segments obtained were combined to generate RTTM per audio resolution. Lastly, we ensemble different multi-resolution diarization systems using DOVER-Lap algorithm. This work was part of DISPLACE challenge organized in INTERSPEECH 2023 and hence the challenge dataset was utilized for all the experiments. It shows that 5-second of audio resolution (i.e.,S-1) yield optimum result of 38.12% and 42.45% DER on development and evaluation data respectively. Furthermore, combining multi-resolution diarization systems (i.e.,S-2) produced an absolute improvement of 3.22% over S-1 and 11.66% over the challenge baseline, with a total DER of 34.9% on the Development set.

Index Terms: Spoken Language Identification, Language Diarization, Audio Accent Identification, Whisper, DOVER-Lap

1. Introduction

Initially, the term diarization was used to identify the audio segments associated with a single speaker in recordings with multiple speakers. It was later expanded to encompass language diarization (LD), a variant of language identification (LI), which involves detecting the languages used in each utterance of a naturally occurring multilingual recording. This study deals with LD in the context of code-switching speech, where a speaker may use several languages, often within the same sentence.

Numerous algorithms have been proposed in the literature for diarization tasks, with the Agglomerative Hierarchical Clustering (AHC) method being widely employed for diarization[1]. AHC starts with an initial speech segmentation in which each segment is assumed to correspond to one language. The AHC strategy involves iteratively grouping these segments until each segment is assigned to its respective language. In each iteration, a pair of clusters are merged, and a new segmentation is created to refine the language turn boundaries. The Bayesian information criterion has conventionally been used to determine which pair of clusters should be merged, whereas Viterbi decoding has been the most common algorithm for re-segmentation. i-vectors or x-vectors have been used to model clusters[2, 3]. Notably, combining the x-vector framework with Probabilistic Linear Discriminant Analysis has shown significant improvements compared to i-vector approaches for clustering. However, this improvement has not yet been demonstrated for the segmentation task.

The literature has seen significant developments in the field of end-to-end (E2E) neural language diarization (LD). Inspired by the success of E2E speaker diarization [4, 5, 6, 7] and the x-vector language encoder [8] researchers have proposed several methods for E2E neural LD. Authors in [9] proposed a neural network-based acoustic-phonetic approach to perform Spoken language Diarization (SLD). From the experimental results it has been seen that the proposed approach is capable of performing SLD tasks when the code-switched utterances have a larger block duration. But the performance of the approach gradually decays with the decrease in the block duration. Similar results were observed in experimentation performed by [10] where the authors proposed a language diarization system that uses combined acoustic and phonotactic cues with the variant lengths of temporal information.

The present study delves into an investigation of the potential applications of the Whisper model decoder, which is a recent development, for the purposes of language identification (LI) and language diarization (LD). The decoder is trained with the aim of predicting a number of features, including language identification, multilingual speech transcription, phrase-level timestamps, and to-English speech translation. In order to improve the LI system's performance, an audio accent detection system is implemented as an attention mechanism to narrow down the output classes. This approach allows for the identification of all accents to a specific detected accent language, which results in a reduction in the number of output languages. The LI system is evaluated using varied audio duration analyses ranging from 1 to 11 seconds. The outcome of the LI system is combined with segment duration in order to produce a final RTTM file for each audio resolution. Multiple multi-resolution diarization systems are then ensembled using the DOVER-Lap algorithm, which has been found to be effective in improving performance in terms of diarization error rate (DER).

The key contributions in this paper are three-fold. First, we propose an attention mask for accent-specific language selection to reduce the false positives for the Language identification task using Whisper decoder. Second multi-resolution windowing method to generate different diarization systems. To the best of our knowledge, the multi-resolution windowing method with Whisper LI has not been studied before for Language diarization systems. Finally, we show that ensembling the various multi-resolution LD system generated using the above steps will improve the performance of diarization systems. The proposed algorithm is not based on clustering rather a detection driven LD. The effectiveness of these methods demonstrates the potential of Whisper-based models for automatic language diarization (LD) task. This paper adds to the existing literature by proposing novel approaches for LD and providing an evaluation of their performance on the challenge dataset. The results of this study will contribute to the development of more accu-

10.21437/Interspeech.2023-1354

rate and efficient LD systems for a range of applications using Whisper.

2. Proposed Approach

2.1. Whisper for language identification

The Whisper architecture is a simple end-to-end approach, implemented as an encoder-decoder Transformer. Input audio is split into 30-second chunks, converted into a log-Mel spectrogram, and then passed into an encoder. A decoder is trained to predict the corresponding text caption, intermixed with special tokens that direct the single model to perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation [11].

2.2. Accent Identification and Language masking

Accents reflect variations in pronunciation, intonation, rhythm, and speech patterns, which arise due to geographical, social, cultural, or linguistic factors. Detecting and understanding accents can provide valuable insights for various applications such as speech recognition, speaker diarization, language learning, and sociolinguistic studies.

Accent detection plays a significant role in language diarization, which involves the identification and segmentation of different languages spoken within a multilingual recording. Integrating accent detection techniques into language diarization systems can provide valuable insights for addressing code-switching challenges. Accents can serve as an additional cue to differentiate between languages and help identify language switches even within code-switching instances [12]. By leveraging accent features along with other acoustic and phonetic information, one can enhance the accuracy of language segmentation and identification, particularly in diverse linguistic contexts [13].

Whisper's *detect_language* function detects the spoken language in the audio, and returns them as a list of strings, along with the ids of the most probable language tokens and the probability distribution over all language tokens. A total of 99 languages is supported by the Whisper model. It is observed that a larger number of languages creates higher confusion resulting in higher false positives for the Language Identification task. To address this, accent specific attention-like mechanism is proposed in this paper that reduces the number of output languages. This research paper presents the implementation of the Emphasized Channel Attention, Propagation, and Aggregation Time Delay Neural Network (ECAPA-TDNN) architecture for the classification of accented speech [14] [15]. Through the utilization of this model, our study focuses on recognizing accents in the CommonAccent dataset, which encompasses a comprehensive range of 21 accent labels [16]. The terms accent identification and accent detection are used interchangeably throughout the paper. The masked vector is proposed by the authors that contain the information of languages spoken in a particular accent (region). This vector was generated manually corresponding to each accent in the classification model. The masked vector $\boldsymbol{m}_i$ has non-zero elements at the positions corresponding to the accent-specific language and zero elements elsewhere. The terms attention mask, language masking, and language selection are used interchangeably throughout the paper.

$$\boldsymbol{m}_i(j) = \begin{cases} 1 & \text{if } j \in L_i \\ 0 & \end{cases}$$

where L_i is the set of languages spoken in a specific accent, i, and j is the index in the array corresponding to each language. The masking operation allows the model to focus on the relevant features of the input vector for language detection based on accent-specific language. We believe that the use of the recently developed Whisper allowed us to demonstrate the effectiveness of our proposed approach in a readily available and easily reproducible manner. It is worth noting that the use of Whisper was chosen based on our review of the literature, which indicated that pretraining of large acoustic models such as Whisper often fails to account for variations in accented speech, resulting in performance degradation for low-resource accents. This limitation aligns well with our proposed pipeline and the integration of an accent detector has resulted in improved performance.

2.3. Proposed Language Diarization

The proposed algorithm utilizes an audio signal as its input and processes it in two different stages. In the first stage, the audio signal is passed through the accent identification model based on which the masked vector gets selected. In the second stage, windowing is applied with a window duration of *w* and an overlapping window duration of *w/2* to the input audio signal. Each frame, s_r is passed through the Whisper's LI system that in turn generates output probabilities vector $\boldsymbol{x} \in \mathbb{R}^{99}$, given by,

$$\boldsymbol{x} = \mathrm{W}(s_r)$$

where W is the Whisper LI model. Medium size Whisper model is utilized for all the experiments performed in this paper. Finally, the Detection vector, $\boldsymbol{D}$ is obtained by applying the element-wise product (Hadamard product) between output probability vector $\boldsymbol{x}$ and masked vector $\boldsymbol{m}_i$, given by,

$$\boldsymbol{D} = (\boldsymbol{x} \circ \boldsymbol{m}_i)$$

To determine and calibrate the language spoken in a particular frame, we use a strategy called a *tolerance-band strategy*. This means that if the difference between the probability scores of the top two languages is less than p_{tol}, and the language spoken in the previous frame is among the two most likely languages for the current frame, then we assume that the language spoken in the previous frame is also spoken in the current frame. We use this strategy to generate the final RTTM after we've collected detection vectors for all frames. For a given audio segment, suppose p_1 and p_2 be the probability scores of the top two languages in the current segment, where $p_1 \geq p_2$. Let l_{prev} be the language spoken in the previous segment, and let l_1 and l_2 be the top two languages in the current segment. Then, we assume that the language spoken in the previous segment is also spoken in the current segment if the following condition holds: $p_1 - p_2 < p_{tol}$ and $l_{prev} \in l_1, l_2$ than $l_{current} = l_{prev}$.

2.4. DOVER-Lap for language diarization (LD) task

DOVER (diarization output voting error reduction) [17] is a weighted majority voting-based method that combines diarization hypotheses. It comprises two different stages, 1. label mapping, and 2. label voting, where each stage addresses one of the problems outlined in the previous section. Similar to DOVER, DOVER-Lap (DOVER + Overlap) is designed to handle overlapping segments in diarization outputs [18] [19]. In this paper, we have explored the DOVER-Lap system for combining multiple diarization systems generated using the multi-resolution (different window duration for audio analysis) setting as shown in Figure 1.

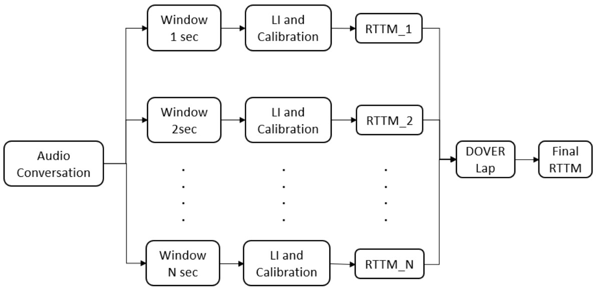

Figure 1: *Proposed multi-resolution windowing method for LD*

3. Experimental Setup

3.1. Dataset

The DISPLACE challenge is a unique task that involves performing both speaker and language diarization on natural, multilingual, and multi-speaker conversations [20]. The conversations, which are around 30 to 60 minutes long, involve 3-5 participants who speak Indian languages along with English (Indian accent). The participants wear close-talking microphones, and there is also a far-field microphone present. The annotations for the data are generated using the close-talking microphones, while the automatic systems will be evaluated on the single-channel far-field audio. The dataset is challenging due to its natural code-mixing, code-switching, reverberation, far-field effects, speaker overlaps, short turns, and pauses, as well as the presence of multiple dialects of the same language. The development dataset consists of 27 conversations with a total duration of 20.6 hours, while the evaluation dataset consists of 20 conversations with a total duration of 11.4 hours.

3.2. Evaluation Metrics

The primary means of assessing the efficacy of speaker diarization systems in the literature is through the diarization error rate (DER). This metric is defined according to the NIST Rich Transcription Spring 2003 Evaluation (RT03S) [21], which outlines the following definition for DER:

$$DER = \frac{D_{FA} + D_{Miss} + D_{Error}}{D_{Total}}$$

where,
D_{FA} = total duration of system speaker segments not attributed to a reference speaker
D_{Miss} = total duration of reference speaker segments not identified by the system output
D_{Error} = total duration of system speaker segments incorrectly attributed to a reference speaker
D_{Total} = total duration of all reference speaker segments

In addition to the above primary metric for evaluation, authors had also evaluated the system on JER, B3-P (BCubed precision), B3-R (BCubed Recall), and B3-F1 (BCubed F1 score) using d-score toolkit [22].

3.3. Effect of Language masking on LD

Firstly, we detect the accent of a given audio conversation using an accent identification system as discussed in section 2.2. In the case of accent identification, a box plot can be used to visualize the output likelihoods for multiple accents for a given audio dataset. Figure 2 and Figure 3 show the box plot for accent distribution of development and evaluation audio dataset respectively. Notably, from Figure 2 and 3, the majority of audios are detected as $Indian$ accent with very high likelihood. It is also observed from the figure that for very few audio files accents are detected as US and $England$ because the spoken language is English. Based on this evidence we are generating the mask m_i by setting the value to 1 for the following languages: English, Bengali, Tamil, Hindi, Kannada, Telugu, Gujarati, Assamese, Malayalam, Marathi, Punjabi, and Urdu (a total of 12 languages out of 99 languages supported by Whisper for LI). To check the effectiveness of language masking, we have fixed the window duration *w* as *5sec* and evaluated the Language diarization system on the development set. We got an absolute improvement of 2.02 % DER compared to without language selection/masking which is 40.14% vs. 38.12%.

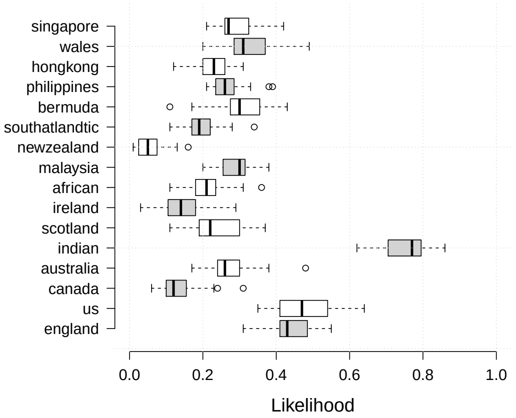

Figure 2: *Box plot representation of accent-wise output likelihood score for all audio files from development dataset*

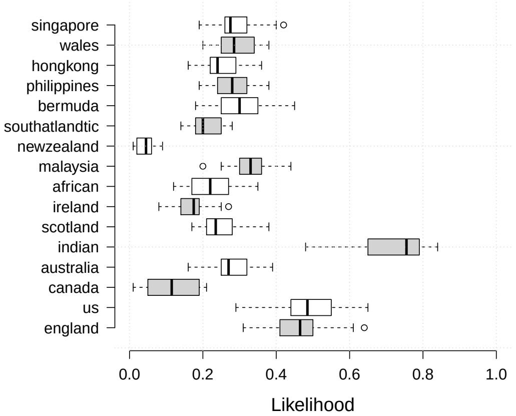

Figure 3: *Box plot representation of accent-wise output likelihood score for all audio files from evaluation dataset*

3.4. Effect of audio analysis window duration on LD

After getting significant improvement in Language diarization (LD) task we did all our further experiments with the language masking strategy only. In this section, we discuss the effect of audio analysis window duration on LD. For that we have varied window length w ranging from *1sec* to *11sec* with 1 sec of linear increment and perform LD as discussed on Section 2. We are getting optimum DER at *5sec* of analysis window duration as shown in Table 1. The same effect was observed with and without tolerance calibration also. It is also evident from Table 1 that p_{tol} = 0.05 further improves DER for all window duration.

Table 1: *Effect of audio window length on DER on Dev set*

Window length **sec**	p_{tol} **100%**	p_{tol} **5%**
1	46.8	45.83
2	42.6	40.99
3	40.68	39.43
4	40	39.05
5	**39.06**	**38.12**
6	39.45	38.64
7	39.12	38.57
8	39.65	39.06
9	39.6	39.16
10	40.24	39.8
11	40.32	39.64

3.5. DOVER-Lap with multi-resolution setting

Further, we have studied the effect of ensembling technique DOVER-Lap on different LD system generated using the multi-resolution windowing method as shown in Figure 1. We will discuss this on next section.

4. Experimental Results

In this section, we discuss the experimental results on the development and evaluation dataset as discussed in section 2. We have fixed the following hyperparameter for all our experiments i.e., w varies from *1sec* to *11sec* with 1 sec of interval, p_{tol} = 0.05.

For a given audio conversation we are generating 11 different LD hypotheses with proposed mask-based language identification and multi-resolution windowing. We have evaluated 3 different systems as mentioned in Table 2.

Table 2: *System descriptions*

System	**Description**
Baseline	Embeddings extraction from pretrained ECAPA TDNN SPEECHBRAIN model and followed by AHC clustering [14][15]
S-1	Proposed Whisper LI with language masking and w = *5sec* audio resolution
S-2	Proposed multi-resolution whisper LID and combination of multiple hypotheses using DOVER-Lap (as shown in Figure 1)

Table 3: *Language Diarization results on Dev set*

System	**DER**	**JER**	**B3-P**	**B3-R**	**B3-F1**
Baseline	46.56	73.44	-	-	-
S-1	38.12	54.48	0.54	0.61	0.57
S-2	34.9	53.01	0.55	0.68	0.61

Table 4: *Language Diarization results on Eval set*

System	**DER**	**JER**	**B3-P**	**B3-R**	**B3-F1**
S-1	42.45	85.58	0.58	0.58	0.58
S-2	36.94	84.95	0.57	0.70	0.63

Table 3 shows the results for the proposed LD system. The *5sec* window duration system (S-1) obtained a DER of 38.12% compared to the baseline 46.56 % on the development set. Notably, multi-resolution analysis window system combined using the DOVER-lap technique system (S-2) obtained a DER of 34.90 % , which is an absolute improvement of 11.66 % over baseline (46.56%) and 3.22 % over S-1 (38.12%). Table 4 shows similar results were obtained for the evaluation dataset with an absolute improvement of 5.51 over S-1 which is 36.94% over 42.45%. We also obtained an absolute of 12 % and 5% improvement for B3-R and B3-F1 score, which indicate the total number of false negative classes reduced significantly for the evaluation dataset.

5. Summary and Conclusions

The newly developed Whisper model decoder has been trained to perform various speech processing tasks such as language identification, speech transcription, and speech translation. This research paper investigates the effectiveness of the Whisper decoder for language identification (LI) and language diarization (LD) tasks. To improve the accuracy of the LI system, an audio accent detection system was used as an attention mechanism to narrow down the output classes from 99 languages to 12 languages. The LI system was tested on audio analysis durations (resolutions) ranging from *1sec* to *11sec*, and the segments obtained were combined to generate the final RTTM file per audio resolution. Lastly, by utilizing the DOVER-Lap algorithm, we ensemble different multi-resolution diarization systems to improve the Diarization Error Rate (DER) performance. This work was part of DIriazation of SPeaker and LAnguage in Conversational Environments (DISPLACE) challenge organized in INTERSPEECH 2023 and hence the challenge dataset was utilized for all the experiments in this paper. The experiment shows that 5sec of audio resolution (i.e., S-1) yield optimum results of 38.12% and 42.45% DER on development and evaluation data respectively. Furthermore, combining multi-resolution diarization systems (i.e., S-2) produces an absolute improvement of 3.22% over S-1 and 11.66% over the challenge baseline, with a total DER of 34.9% on the development set. Similarly, S-2 gave an absolute improvement of 5.51% over S-1 on the evaluation set. Our results demonstrate the potential of the Whisper model decoder for language diarization applications. The authors would also like to emphasize that the primary novelty of the paper lies in the pipeline approach that we proposed for language diarization. Specifically, we utilized an accent detector along with a language identification system to obtain multiresolution language diarization outputs.

6. References

[1] H. Aronowitz, W. Zhu, M. Suzuki, G. Kurata, and R. Hoory, "New advances in speaker diarization." in *INTERSPEECH*, 2020, pp. 279–283.

[2] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 413–417.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[4] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[5] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 296–303.

[6] T. Park, M. Kumar, and S. Narayanan, "Multi-scale speaker diarization with neural affinity score fusion," in *In proceedings of ICASSP*, May 2021.

[7] D. T. Toledano, M. P. Fernández-Gallego, and A. Lozano-Diez, "Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on timit," *PLOS ONE*, vol. 13, pp. 1–24, 10 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0205355

[8] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors." in *Odyssey*, vol. 2018, 2018, pp. 105–111.

[9] J. Mishra, A. Agarwal, and S. R. M. Prasanna, "Spoken language diarization using an attention based neural network," in *2021 National Conference on Communications (NCC)*, 2021, pp. 1–6.

[10] L. Dau-Cheng, C. Eng-Siong, and L. Haizhou, "Language diarization for code-switch conversational speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7314–7318.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[12] A. Siddhant, P. Jyothi, and S. Ganapathy, "Leveraging native language speech for accent identification using deep siamese networks," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 621–628.

[13] K. Kukk and T. Alumäe, "Improving Language Identification of Accented Speech," in *Proc. INTERSPEECH 2022*, 2022, pp. 1288–1292.

[14] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *INTERSPEECH 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[15] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[16] Juan Pablo Zuluaga, "Accent-ID-commonaccent-ECAPA," https://huggingface.co/Jzuluaga/accent-id-commonaccent_ecapa, Last Accessed: 05 March 2023 [Online].

[17] A. Stolcke and T. Yoshioka, "Dover: A method for combining diarization outputs," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 757–763, 2019.

[18] D. Raj, P.Garcia, Z.Huang, S.Watanabe, D.Povey, A.Stolcke, and S.Khudanpur, "DOVER-Lap: A method for combining overlap-aware diarization outputs," *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[19] D. Raj and S. Khudanpur, "Reformulating DOVER-Lap label mapping as a graph partitioning problem," *INTERSPEECH*, 2021.

[20] S. Baghel, S. Ramoji, Sidharth, R. H, P. Singh, S. Jain, P. R. Chowdhuri, K. Kulkarni, S. Padhi, D. Vijayasenan, and S. Ganapathy, "DISPLACE Challenge: DIarization of SPeaker and LAnguage in Conversational Environments," 2023.

[21] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre, "The elisa consortium approaches in broadcast news speaker segmentation during the nist 2003 rich transcription evaluation," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–373.

[22] N. Ryant., "dscore: Diarization scoring toolkit," Jan. 2019. [Online]. Available: https://github.com/nryant/dscore